DOCUMENT RESUME

ED 292 853 TM 011 233

AUTHOR Burt, Charles W.
TITLE An Experimental Study of Group Size and Participants' Role in Developmental Testing.
PUB DATE Apr 87
GRANT FCAR-290-10
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). Some tables have poor type.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Foreign Countries; *Formative Evaluation; *Group Testing; *Instructional Materials; Nursing Education; Research Methodology; Student Attitudes; Two Year Colleges
IDENTIFIERS *Group Interviews

ABSTRACT

The research in progress that is presented in this paper focuses on the preliminary results of a study examining the differences and similarities in data obtained using the one-to-one versus small group procedures for designing and conducting formative evaluations of instructional materials. The materials chosen were the microbiology print module of the "Instructional Materials for the Dental Health Profession." When selected for the study, the materials were in draft form, needed revision, required 1 hour of learner time, concerned a subject unfamiliar to the target population, were developed by instructional designers applying guidelines of systematic instructional development, and included criterion-referenced tests. The materials were chosen for first-year nursing students at two junior college campuses offering nursing programmes in English on the Island of Montreal. Thirty-six subjects on each campus were assigned to 1 of 4 one-to-one or small group conditions. Data were gathered in via a posttest, attitudinal questionnaire, and debriefing sessions. Results indicate that although there are few significant differences in the amounts of data gathered by the techniques under investigation, different techniques do result in different kinds of feedback data. An appendix contains numerous data tables and figures. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED292853

An Experimental Study of Group Size and Participants' Role in Developmental Testing

Charles W. Burt

McGill University

A paper presented at the annual meeting
of the American Educational Research Association,
Washington D.C., April 1987

Preparation of this paper was supported by the Quebec Programme de formation de chercheurs et aide a la recherche (FCAR), Grant #290-10

TM 011 233

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles W. Burt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

There exists a plethora of instructional design models in the literature from which to choose. A major review of existing models by Andrews and Goodson (1980) compared 40 models in terms of their differences and similarities across components (procedures) of the design process. This paper will focus on one of the components of the design process: designing and conducting formative evaluation. Andrews and Goodson found that 38 out of the 40 reviewed models endorsed formative evaluation as a process through which data can be collected from members of the target population in order to revise and improve the effectiveness of a particular instructional design effort while it is still being developed. Similar follow-up reviews of design models by Stolovitch (1982) and Geis, Burt, and Weston (1984) concur that formative evaluation is a well-endorsed procedure. Burt and Geis (1986) reported the results of a survey of instructional design practitioners which concluded that formative evaluation was being conducted by all of the respondents to their survey.

A basic rationale for doing formative evaluation appears to be related to the nature of design models. Models are presented as design heuristics. Applying a heuristic does not guarantee that the design goals/objectives will be attained. For this reason, it is prudent to review and revise an instructional design effort in order to monitor and improve the resulting instructional product. This should be done while the product is in "rough draft" form as opposed to a more polished version. Research by Rodriguez and Rodriguez (1982) indicates that when formative evaluation is applied to audio-visual learning materials at three different levels of product development (story-board, unprofessional videotape, and professional16mm film) there are no significant differences in the feedback that students provide to the developer/evaluator. They recommend that if tryout is to be done only once during the production, it should be conducted very early in the development of the product as feedback will be effective at this stage. Modifications done at the early stage will cost less than at later stages of development.

Formative evaluation can be conceptualized as a continuum of procedures with different variations or approaches being endorsed by different authors within the literature. Weston (1986) reviews all of the different approaches being endorsed for the evaluation of instructional

materials. Two basic distinctions can be made within formative evaluation approaches. One approach suggests that review by instructional design experts is sufficient. For example, Montague, Ellis, & Wulfeck (1983) have conducted research on the Instructional Quality Inventory (IQI) which is analogous to a checklisting tool that requires the developers to recycle back through all of the components of the design process in order to look for inadequacies and misalignments of design decisions and implementations. Another, more widely endorsed approach, recommends combining the use of appropriate experts (e.g., measurement/evaluation experts, subject matter experts, media/communications experts, developmental psychologists) as consultants within their particular domains of expertise in conjunction with empirical tryout data from members of the target population for whom the material was designed (i.e., student test subjects). This approach, for example, recognizes that students are unable to comment on the validity of the instructional content, but that they might be able to provide useful feedback with regards to the problems they are having while trying to learn from the instruction and they can report on their affective reactions to the learning experience. The combined approach of expert review and collecting tryout data also recognizes, through the research of Newman (1957); Rothkopf (1963); Rosen (1968); Frase, DeGracie & Poston (1974); Henderson and Nathenson (1977); and Smith (1978) that experts are not always successful in predicting what kinds of problems will be encountered by learners (i.e., predicting the effectiveness of the instruction). Within this combined approach, six distinct variations of collecting empirical tryout data from student test subjects have been identified within the instructional design literature:

1. One-to-One
2. Small group
3. One-to-One plus small group
4. Field testing
5. 3 stage model (one-to-one plus small group plus field testing)
6. Learner verification and revision (LVR)

The research in progress that is being presented in this paper will focus on the preliminary results of a study examining the differences and similarities of data obtained using the one-to-one

and/or small group procedure for designing and conducting formative evaluations. These procedures are included within all of the above listed variations of formative evaluation with the exception of field testing which uses larger groups of student test subjects. In order to maintain a distinction between these variants of formative evaluation, the term developmental testing will be used to include one-to-one and/or small group formative evaluation. The early work of Horn (1966) and others researching programmed instruction serves as a historical precedent for coining the term "developmental testing". This is still a useful descriptor as it emphasizes obtaining feedback data while the instructional product (often self-instructional print materials) is still being developed.

Within the research literature related to developmental testing, the emphasis has been on comparing the effectiveness of instruction revised based on either one-to-one sessions, small group sessions, or on both one-to-one and small group sessions. This research strategy appears to address only one of Baker and Alkin's (1973) research questions related to formative evaluation: "1. What are the major types of data to obtain and the major sources of such data? and 2. What is the impact of evaluation data upon the modification of programs?" (p. 393). The effectiveness of revised versions of instructional material under varying experimental conditions is a frequent dependent variable within the developmental testing of instructional materials research literature because the reason, in practice, for conducting testing sessions is to obtain data for purposes of revision. Although research, to this point, has tended to emphasize the relative effectiveness of revisions, the importance of the sources and kinds of data produced from them should not be ignored. It is quite feasible in supposing that the data produced under one-to-one and small group conditions will differ, otherwise why bother conducting product tryouts using both strategies. Burt and Geis (1986) suggest that, within one-to-one/small group sessions, evaluators utilize a number of different sources for collecting feedback data. Student test subjects articulating their problems while trying to learn from the material, posttest errors, affective reactions to a questionnaire, and debriefing results after the test session all can provide data. Differing roles that the participants could play during a testing session (e.g., an actively intervening and probing

evaluator versus a more passive observer) might affect the data from these sources. However, no research until now has systematically described the data from these different sources and conditions. A better understanding of these data utilizing various formative evaluation methodologies should ultimately facilitate our understanding of the revision process — but this research does not look at the process of revision nor the effectiveness of revision. Burkholder (1982), reviewing the literature since 1960, views the research related to the revision of instructional materials in three particular areas:

1. revision methodology
2. case studies explaining the revisions made; and
3. the derivation of principles for revising instruction

Dick (1980) has been consistently monitoring and contributing to the research on formative evaluation and notes that:

> The state-of-the-art in formative evaluation at the present reflects a relatively mechanistic approach to the process. . . What is needed is a more theoretical approach which would a) suggest the kinds of data which should be collected and b) suggest how the data should be interpreted by the designer and used to revise the instruction. Such an approach would indicate the extent to which the instruction reflects the theory and is instructionally effective and the ways in which the instruction should be improved to better reflect the theory (pp. 5-6).

In one of the few studies employing a revision algorithm, Dupont and Stolovitch (1983) compared expert review with Gropper's (1975) non-validated theoretical revision model which utilizes test subject data. Results indicated that materials revised using the revision model and test subject data did not produce significantly superior attainment of the objectives when compared to using only test subject data. Also, materials revised on the basis of student data produced significantly superior attainment of learning objectives on an achievement test on the

manipulation of an SLR camera, but not on a paper and pencil test on the principles underlying the use of an SLR camera when compared to expert (subject matter expert) revision.

In an early and frequently cited study Robeck (1965) addressed the feasibility of using a single student (one-to-one) to provide the database upon which revision was to be grounded. A revised version was produced based upon the reactions of one test subject (a bright male grade six student) to the prototype. Another revised version was produced based on the reactions of another individual test subject. The three versions of the material (the prototype, version one, version two) were then randomly assigned to three matched groups of students. The two revised versions significantly outperformed the prototype. There was no significant difference between the performance of versions one and two. This study does demonstrate that even revising based solely on a single student can significantly improve instruction. The only problem with adopting this practice is how to select an "appropriate student"?

Wager (1983) used Dick and Carey's (1978) procedures for conducting both one-to-one and small group tryouts in an experiment addressing both "the appropriate student" and the number of subjects issues. From pools of high, medium, and low aptitude groups, it was found that high aptitude subjects in the one-to-one condition were more adept at locating inaccuracies and generating instruction for a self-instructional print module when it was perceived as deficient. Low aptitide students identified the more basic problems such as the use of inappropriate vocabulary within the module. A mixed aptitude pool, providing feedback during one-to-one, produced the greatest variety compared to homogeneous aptitude groups. Students who received materials revised based on the feedback from the mixed aptitude group scored significantly higher on the posttest and attained the most favorable attitude rating than did students who used materials revised either on the basis of feedback from the high or low aptitude group. Wager's study also indicated that materials revised based on feedback from the one-to-one sessions using subjects from the mixed aptitude group were as effective as materials revised based on a combination of one-to-one and small group sessions. The materials revised based on one-to-one sessions from the mixed aptitude group were more effective than materials revised according to feedback from a

small group session alone. Berthelot (1978) was also interested in test subject characteristics as an important variable during one-to-one data collection. He compared high, medium, and low levels of test subject verbal ability with revison outcomes. Although there were no significant differences in revisions, best results were obtained from the highly verbal subject.

Kandaswamy, Stolovitch, and Thiagarajan (1976) conducted a study to compare the effectiveness of one-to-one and small group-based LVR. One-to-one and group conditions used equal proportions of "high", "medium", and "low" level students. In the one-to-one condition, subjects provided reactions, responses, and remarks to the evaluator/reviser, who did not instruct students when difficulties were encountered but did probe when difficulties were observed. Group-based LVR subjects only provided pre/posttest data and within program response data — no debriefing data were collected. All versions of the instructional material revised under these conditions were administered to groups of students to determine their relative effectiveness. A Scheffe multiple contrast test demonstrated that all revised versions were more effective than the original. There were no significant differences between group based and one-to-one LVR. A number of evaluators/revisers were utilized in the study and this was a factor in the experimental design. An evaluator/reviser main effect was observed indicating that different evaluators/revisers produced varying degrees of effectiveness for the materials.

In a similar study, Baghdadi (1980) was interested in determining: if revised materials were better than unrevised materials; if there were any differences between one-to-one and small group LVR approaches; and if different revisers produced varying degrees of effective instructional materials. The instructional material used in the study was a programmed unit designed for graduate students on four different evaluation models. Four evaluators familiar with LVR procedures were hired to revise the materials. Two of these evaluators independently conducted one-to-one evaluations while the other two conducted small group sessions. The two groups of evaluators then exchanged the methods to produce four other versions of the material. The eight revised versions and the original version were randomly assigned to target learners. Revised versions were significantly better in attaining design objectives than the unrevised

version. No significant differences were detected between one-to-one and small group approaches. No reviser effect, in terms of producing revised materials of varying effectiveness, was found in this particular study.

The studies reviewed indicate that formative evaluation does produce a more effective product than the original unrevised draft version. Debert (1979) has produced a classification scheme for the general kinds of revisions that instructional developers tend to make based on the problems encountered by test subjects during materials tryouts. A simple inspection of the categories demonstrates that subjects are able to comment on problems encountered with materials at a level that reflects a broad spectrum of the kinds of decisions that are made in order to produce instruction. However, as Weston (1987) notes, the revision process can be viewed as mainly subjective or intuitive as there are few guidelines to aid a designer in translating developmental testing data into effective revisions. A major goal of formative evaluation research is to produce such validated guidelines. Through the accumulating research literature, there are examples of such efforts. A case in point is the work of Dupont (1980) who has produced an algorithm that facilitates the translation of the data collected into the kinds of revision prescriptions identified by Debert (1979).

The reviewed literature has still not reached concensus on the relative effectiveness of one-to-one and/or small group testing — both produce more effective instruction than the original. This is reflected in the results of the survey conducted by Burt and Geis (1986) where no significant differences were found between proportions of developers using only small group procedures versus both one-to-one and small group procedures in combination. What is clear from the literature is that revised versions are more effective than unrevised versions. It is also clear that the exclusive use of expert evaluation is not enough. The characteristics of optimal test subjects still remains unclear beyond the common-sense recommendation of selecting "representative" subjects from the target population. The research literature to date has not focused on the kinds of data collected during developmental testing; rather it has concentrated on revision outcomes. In addition, the instructional design textbook literature as reviewed by Geis, Burt, & Weston (1984)

presents conflicting guidelines as to the kinds of roles that the participants of testing sessions (student test subjects and designer/evaluator) should play. The reviewed research has not addressed this important variable. For these reasons, the following study was conducted. The results of the study are based on an anaylsis of the responses that student test subjects provide under different developmental testing conditions. The study does not deal with the effectiveness issue although there may be implications for effectiveness. The experiment reported here was conducted for descriptive purposes. The intent of this research is to describe the amounts and kinds of information (feedback) developmental testing provides, as a process utilizing a number of data collection methodologies.

METHODOLOGY

Instructional Materials Description

The learning materials used in all experimental developmental testing sessions consisted of the first two units/chapters of the print module "Microbiology", one book in the series: Instructional Materials for the Dental Health Professions (Dental Auxiliary Education (D. A. E.) Project, 1982). These materials were available in rough (first) draft form and were selected because they:

1. were in draft form
2. were less than "perfect" (i.e., need revision)
3. were in replicable rather than ephemeral form (e.g., print rather than lecture)
4. required about an hour of learner time
5. concerned a subject matter that was unfamiliar to many people in the target population
6. did not require prerequisite knowledge that the target population would not have
7. had been developed by instructional designers applying the guidelines of systematic instructional development
8. provided criterion-referenced tests developed by the authors

Materials description. The Microbiology module is only one component of a series of self-instructional modules that comprise a major part of a complete dental auxillary curriculum designed to train professional dental assistants and dental hygienists. As described in the Faculty Guide for the complete set of curriculum materials, the Microbiology module is considered to be a

sub-module within the curriculum because it covers clinical, basic science, or dental science information that is needed throughout the whole training program.

The materials were developed using teams of content experts, instructional designers, and evaluators. A modification of Horn's (1976) Information Mapping © system served as the basis for formatting the modules according to the subject matter and characteristics of the target population. Using this system resulted in the presentation of the information in blocks which are separated by lines in order to facillitate the efficient location of information. Beside each block of information, marginal labels were used to identify the subject and type of information. This was intended to aid the reader scan the module for the purposes of locating or reviewing important sections of information. Illustrations were used to aid in clarifying the printed material. Introductions are used at the beginning of each unit to preview the reason for learning the material. The objectives for each unit/chapter of any given module were presented to the learners in a detailed and behaviorally concrete manner following the introduction of each chapter. The grouping and sequencing of units within the modules were determined by the type of information being conveyed. The two units used in the experiment would be classified as scientific and conceptual information. With this kind of information, a separate section or block is used to introduce new terms, topics, or conceptual information in general. Whenever there were categorical schemes to be learned, learners were presented separate sections in a list format that explained relationships, diagrams, and outlines.

Sample

Of critical concern to this study was the matching of the materials with an appropriate target population. The materials had been originally designed for dental hygiene students but it was discovered that there were too few dental hygiene programmes (offering instruction in the English language) from which to select a suitable number of test subjects. A curriculum development specialist at McGill University, familiar with the college-level nursing curriculum, reviewed the materials and suggested that the materials would suit first year nursing students. Of the three

junior colleges offering nursing programmes in English on the Island of Montreal, two were randomly selected. Both the chairpersons and academic teaching staff of these two programmes reviewed the instructional materials and judged them to be suitable for the first year students. None of these students had been exposed to Microbiology within their programmmes as this was a course taken in the second year of study.

Two campuses were canvassed for volunteers. Test subjects were informed of the nature of the study and were told that they would each receive a $50.00 stipend for their time. From each college thirty-six test subjects were randomly selected from the volunteer pool. A total of seventy-two subjects were utilized in the study.

Study Procedures

From each group of thirty-six test subjects, subjects were randomly assigned to one of four experimental conditions: 1) one-to-one active; 2) one-to-one passive; 3) small group active; or 4) small group passive. Eighteen one-to-one active sessions (N=18), eighteen one-to-one passive sessions (N=18), three small group active sessions with 6 subjects per small group (N=18), and three small group passive sessions with 6 subjects per small group (N=18) were conducted by the same experimenter.

All conditions. For both the one-to-one and small group experimental conditions, subjects were instructed that the materials were in first draft form and that the purpose of developmental testing was not to test how well subjects could learn, but to learn how well the materials taught. All subjects were told that their feedback on the effectiveness of the material was important. Test subjects were asked to locate the problems they were encountering (any kind of problem), to explain the nature of the perceived difficulty, and to suggest, if possible, a revision addressing the problem so that another student would not encounter the same problem. For each condition subjects were given specific instructions for purposes of task orientation. All subjects were asked to leaf through the instructional material, while the experimenter instructed them to attend to the posttest and the attitudinal questionnaire. They were then told that there would be a short

break after which they would be debriefed (interviewed) based on their responses to the material. A tape recorder was used to record the interactions between the test subject(s) and the experimenter.

One-to-one active. During active one-to-one developmental testing, each subject was instructed to go through the materials as best they could and if either the subject or the experimenter perceived that there was a problem with the materials it would be discussed during the session. Cues or signals indicating that a problem was perceived included either verbal behavior: the test subject making some kind of a spontaneous comment or a subject's response to the experimenter's page-by-page probe: "Were there any problems on that page?" ; or non-verbal test subject behavior signalling to the experimenter that there might be a problem (i.e., rapid page-turning, fidgeting, or when the subject started highlighting/writing on the material). Subjects were given a yellow highlighting marker and instructed to use it in order to identify/locate their perceived problems by underlining the problematic portion(s) of the textual display. If a test subject indicated that there was a problem or if the experimenter perceived there to be a problem, the experimenter responded by appropriately saying to the individual test subject either:

1. "What is the problem?" **or**

2. "Subject's name, I've noticed that you have been (e.g., fidgeting). Are you having a problem?"

Under either circumstance, the experimenter probed the test subject until there was a definite indication of a problem. If a test subject persisted in asking content questions of the experimenter, he would respond by saying: "Where were you in the material when you began to feel that you had a problem?" The experimenter did not provide instructional assistance, only probing whenever difficulties were observed. In all circumstances the subject would locate the problem area resulting in either the experimenter or the subject underlining the portion(s) of the text with the highlighter. Then the experimenter probed further by saying: "What is the problem?" Thus, before going on, the problem was explicated by the test subject resulting in either the subject or the experimenter writing on the text the nature of the perceived problem. Upon explication of the problem, the test subject was asked to suggest a way of avoiding that

problem for the next student who would encounter the instructional material. The experimenter, in order to operationalize a revision "set", responded by saying: "What could be changed within the material so that somebody else wouldn't have that problem?" If the subject was unable to generate a revision then he/she was asked to move on within the instructional material. These experimental "instructions" were written on a cover sheet for the materials and were verbally explained at the onset of the experiment.

One-to-one passive. During passive developmental testing, subjects were instructed to go through the materials as best they could, use the highlighter to locate any problem area(s), use a pen provided to explain the problem by writing of the material itself, and finally, if possible, to make/suggest a revision. After explaining the task for the subject, the experimenter left the room in order to let the subject work on their own.

Small group active. During small group active developmental testing, subjects were given identical instructions as described for one-to-one active. The only differences between the conditions were that subjects were requested not to talk with their "neighbors" during the session as there were now six test subjects sitting around a large table with the experimenter. Subjects were requested to raise their hand if they wished to express that they perceived a problem.

Small group passive. For small group passive developmental testing, subjects were given identical instructions as described for one-to-one passive. Similarly to small group active, subjects were requested not to talk with their "neighbors" during the session. The experimenter did stay in the room with these subjects but did not interact or probe subjects as difficulties arose. These subjects were individually on their own locating problems with the highlighter, explaining the problems on the materials in writing, and if possible, making/suggesting revisions in writing.

Instrumentation and Data Sources

The experimental design utilized in this study called for the testing of an individual subject or a group of subjects over a consecutive two-day period. On day one, the first unit of the material was developmentally tested. On the second day, the second unit of the material was tested.

In addition to test subjects identifying/locating perceived problems, explicating why they perceived something to be problematic and suggesting ways in which that problem could be remedied, they were also requested to complete a criterion posttest (they were permitted to criticize the posttest as well), and complete an attitudinal questionnaire. The criterion tests were designed based on the objectives expressed within the material and the tests sampled the attainment of specific objectives for which subjects had received instruction. Items on the tests sampled a subject's recall of the content. Higher levels of comprehension/performance were not tested. The test for unit one contained five items using both multiple-choice and fill-in-the-blank formats. The test for unit two contained 12 items (reflecting a longer unit than the first) using multiple-choice, fill-in-the-blank, and short-answer formats.

An adapted version of Abedor's (1972) attitudinal questionnaire was utilized in this study. The original questionnaire had been designed for use within a multi-media instructional delivery system, i.e., a self-instructional print workbook in combination with audio-visual media. The adapted version reflected the use of a self-instructional print module as the medium of instructional delivery. The same adapted version was used over the two instructional units. The adapted version contained 27 items of which 23 were on a five point Likert-type scale ranging from strongly agree to strongly disagree. Four items were of open-ended format. Overall, the items were general in nature, requiring a subject to globally evaluate the instructional material. Statements on which students would base their ratings (the 23 Likert-type scaled items), reflected the major decisions that are made during an instructional design effort using a systematic design model.

After having completed a unit of the materials, including the posttest and the attitudinal questionnaire, the subject(s) were given a 15-20 minute break. During this time the experimenter compiled an agenda for debriefing. Criteria for an item appearing on the debriefing agenda varied depending on experimental condition. For one-to-one conditions, all problems identified (test subject comments) while going through the material including comments made while writing the posttest, all errors made on the posttest, any response to a Likert-type scaled

item that was two points away from the ideal response and unfavorable open-ended questionnaire responses were put on the agenda for discussion during debriefing. For small group conditions, if two or more test subjects perceived a problem within the same section of the instruction (while going through the materials both trying to learn and criticize simultaneously) then it was placed on the agenda for debriefing. If two or more test subjects made an error on the same posttest item, the item and the section of the material which provided instruction for that particular item became an agenda item for debriefing. If the groups' mean response to an attitudinal questionnaire item was two or more points away from the ideal response, that item was discussed during debriefing. A similar procedure was applied to the open-ended questions — if two or more subjects responded unfavorably, it became a debriefing agenda item.

After the break, the experimenter debriefed the test subject(s) using a tape recorder to record the session. Debriefing can be conceptualized as a retrospective data collection procedure used to finalize the outcomes of developmental testing for a given individual test subject or group of test subjects. After debriefing, the test subject(s) was asked to reconvene for the next day to developmentally test the second unit of the materials.

EXPERIMENTAL DESIGN AND RESEARCH QUESTIONS

Across all experimental conditions, two kinds of data were collected: qualitative and quantitative. Figure 1 presents a two-factor fully crossed ANOVA experimental design utilized to describe the data collected under one-to-one and small group conditions (number of participants) and role of participants conditions (active and passive). This paper will focus on the results of the quantitative data analysis and present a content analytic framework based on the qualitative data. The following descriptive research questions will be addressed in this paper:

1a. Do experimental conditions affect the amount of data obtained from the different sources of feedback?

1b. Do experimental conditions affect the kind of qualitative data obtained from the different sources of feedback?

2. Does the debriefing procedure produce unique feedback data?

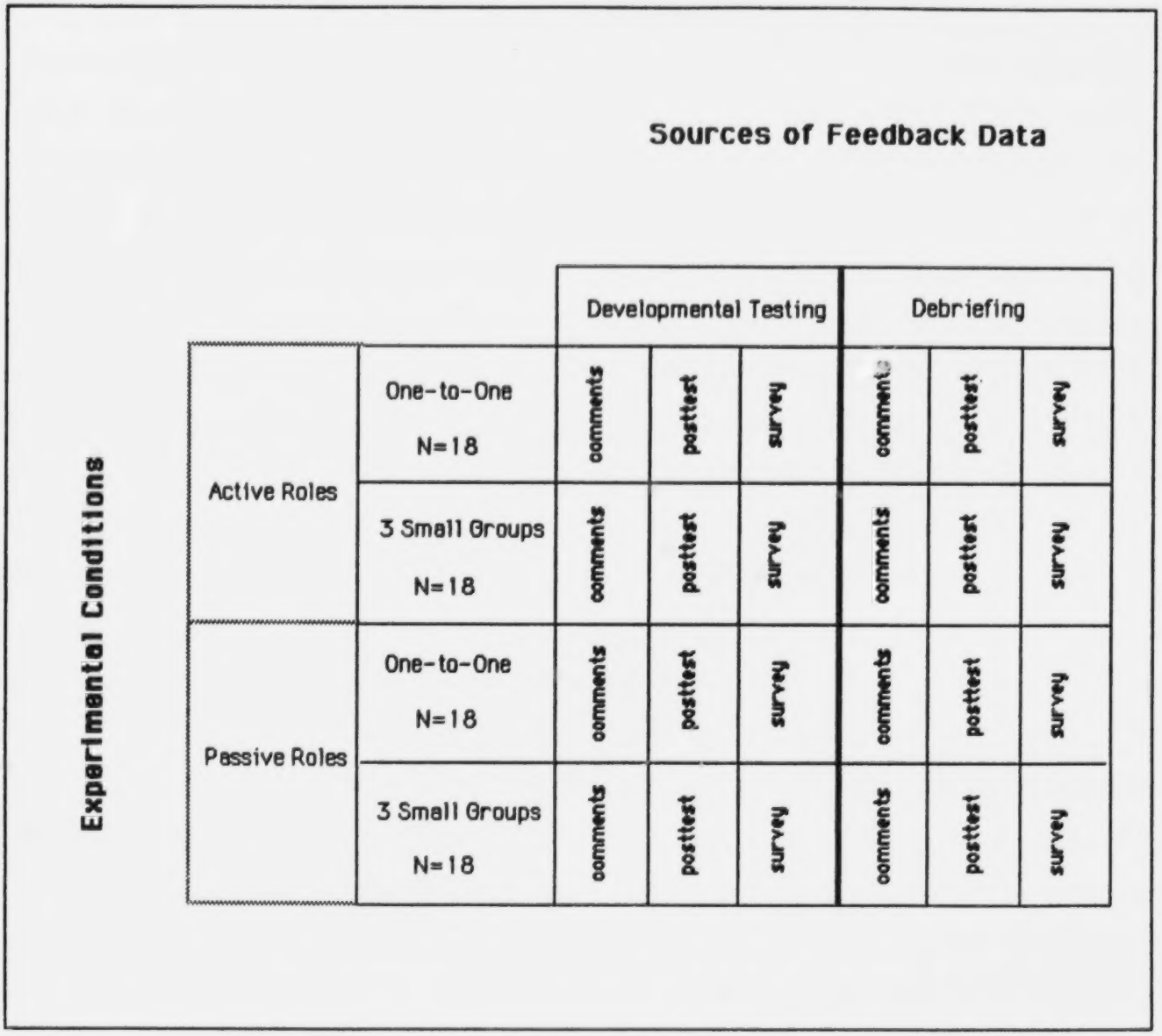

Figure 1. The experimental design: A two-factor fully crossed ANOVA of quantitative data (N=18 per cell) and a content analysis of qualitative data (N=18 per cell). A total of N = 72 subjects participated in the study over two days (two units of instructional material). One subject was lost for the second day.

3. Does the test subject characteristic of having indicated prior knowledge of the material affect the overall amount of comments being generated during developmental testing?

RESULTS

Data analysis procedures were conducted using SPSSX Version 2.1 on an AMDAHL mainframe computer. For the first research question investigating the amount of data obtained from the experimental conditions, it was necessary to conduct separate univariate analyses by the two independent variables: role of participants and number of participants, for the different sources of feedback data (dependent variables): a) the number of comments (problems being identified) made while going through the instructional materials - i.e., developmental testing, b) the number of comments that were debriefed, c) the number of debriefing agenda items based on posttest errors, d) the number of debriefing agenda items based on the attitudinal survey responses (unfavorable responses), and e) the total number of debriefing agenda items from all sources combined across the two units that were developmentally tested experimentally.

Table 1 (see Appendix 1) presents the 2-way ANOVA summary table of the main effects and interactions of independent variables on the number of comments (problems being identified) made while going through the instructional materials - i.e., developmental testing.. Note that separate analyses were conducted for unit 1, unit 2, and for the total (unit 1 + unit 2). Figures 2 and 3 (see Appendix 1) present a graphic portrayal of the results of the data anaysis. The only significant result observed was the main effect for number of participants, $F(1,68) = 4.68$, $p < .05$; one-to-one produced more comments than small group for unit 1.

Table 2 presents the 2-way ANOVA summary table of the main effects and interactions of independent variables on the number of comments that were debriefed. Figures 4 and 5 graphically portray the data. This analysis was of interest because a criterion was applied to the small group condition. At least two subjects had to have identified a problem (make a comment) within the same section of the instructional material in order for a problem to appear on the debriefing agenda. The one-to-one condition did not have such a criterion operating. In all experimental conditions, with respects to debriefing test subject comments, subjects were

permitted to identify problems that had not been noticed during developmental testing. Subjects were also free to change their minds — what was a problem during developmental testing might not have been perceived as problematic upon debriefing and therefore was dropped from the debriefing agenda. The only significant result obtained was the number of participants main effect for the first unit, $F(1,68)= 7.16$, $p< .05$, producing more comments than small group for unit 1.

Table 3 presents the 2-way ANOVA summary table of the main effects and interactions of independent variables on the number of debriefing agenda items based on posttest errors. The data has been plotted in Figures 6 and 7. Similarly to debriefed comments, there was a criterion applied to small group test results. At the individual item level, at least two subjects had to have gotten the item wrong in order for the item to be brought to the debriefing agenda. A significant interaction between role and number of participants was found for unit 1, $F(1,68)= 12.44$, $p< .05$, for unit 2, $F(1,67)= 4.01$, $p< .05$, and for units 1 and 2 combined, $F(1,68)= 11.274$, $p < .05$.

Table 4 presents the 2-way ANOVA summary table of the main effects and interactions of independent variables on the number of debriefing agenda items based on unfavorable attitudinal survey responses. Figures 8 and 9 graphically display the data. Again, there was a criterion applied to the small group condition. The groups' mean response was calculated for each Likert-type scaled item. If the mean response was 2 or more points away from the ideal response it was brought to the agenda. For the open-ended formatted survey items, at least two subjects had to have responded negatively in the group condition. With the one-to-one condition, any response that was two or more points way from the ideal was brought to agenda along with any unfavorable open-ended responses. No significant interactions were found but there were significant main effects for unit 2 and the total of units 1 and 2. With unit 2, both the main effects of participants' role, $F(1,67)= 7.26$, $p< .05$, and number of participants, $F(1,67)= 7.87$, $p< .05$, were statistically significant. For the total number of debriefed survey items across the two units, only the number of participants produced a significant main effect, $F(1,68)= 6.446$, $p< .05$.

Table 5 presents the 2-way ANOVA summary table of the main effects and interactions of independent variables on the total number of debriefing agenda items from all sources combined (comments, posttest results, and survey results) across the two units. Figure 10 is a graphic representation of the data. No significant differences were detected.

DISCUSSION

Usually the inferential statistic of ANOVA is associated with hypotheses testing. In this study there were no research hypotheses — there were descriptive research questions guiding the investigation. There are also methodological problems in interpreting the above results based on the experimental design. It is difficult to compare the results of unit 1 and unit 2. They could be conceived of as two different experiments. The results of units 1 and 2 are contaminated by a number of uncontrolled factors. Although the experimental procedures were the same for both units, unit 2 was a much longer lesson that the first. The length of the second posttest reflects this. Differences in response frequencies by test subjects across the two units could be attributable, for example, to there being inherently more problems in one unit than the other — this is not known. Or, there might be a group dynamics effect or peer effect between the two units (tested over two days) because subjects become less inhibited in making responses after becoming more familiar with each other and the experimenter. The empirical (c.f., McGrath & Altman, 1966) and theoretical literature (c.f., Corey & Corey, 1982) on group processes suggests there are phases through which small groups are thought to develop. Possibly subjects are more aware of what is expected of them (a training effect) during developmental testing and debriefing for the second day. One can only speculate on possible contaminating factors.

In light of the above discussion, the author will concentrate discussion over the two days, i.e., to look at the total results of unit 1 and 2 combined using the ANOVA data. Overall, there are no differences between the active and passive roles for generating comments. Similarly, there are no differences between the one-to-one and group procedure for generating comments (although there was a significant main effect during unit 1 where the one-to-one procedure did produce more

comments). The same can be said for the amount of comments debriefed. With respects to debriefing posttest errors, there were consistently more errors debriefed from the small group passive procedure. In speculation, this was the most restrictive of all of the experimental conditions in the sense that the students were using the materials in a manner which would least resemble their actual use. For the number of survey items debriefed, the small group passive condition appears to produce more overall unfavorable feedback data. When responses are pooled across all feedback sources for the two units combined, there are no significant differences for experimental conditions.

The research being conducted is also looking at the different kinds of data obtained under the different experimental conditions for developmental testing and debriefing. The different kinds of problems being explicated and revisions made (both were considered seperately as test subject comments) by using the various sources for collecting data are qualitative responses made by individual or groups of subjects. To date a content analysis has been conducted across sources over the two days while ignoring the experimental conditions for the total number of test subjects who participated in the study (N=72). The result of this has been the production of a content analytic framework for inspecting the qualitative data. This classification scheme was based, in part, on Debert's (1979) list of revision prescriptions. It was found to be inadequate in accommodating the qualitative data generated by the experiment being reported on by this author. The following is a list of the major categories of the expanded Debert (1979) revision prescription framework. Some examples of the subcategories for each major category are provided but by no means are they presented as being exhaustive of the complete set of subcategories:

1. Revision based on comprehension/recall problems, e.g.:
   - add illustration
   - rearrange the sequence
   - delete an example

2. Revision based on motivation/attention problems, e.g.:
   - increase level of complexity
   - add illustration
   - delete redundancy

3. Revision based on physical features/layout problems, e.g.:
   - use point form (list)

- underline important terms
- change the typeface

4. Revision based on responses overlapping two or more of the above categories, e.g.:
   - delete redundancy + rearrange sequence
   - highlight using capital letters + add explanation
   - add an example(s) + add illustration

5. Non-classifiable responses, e.g.:
   - make content accurate
   - add another response to be considered as correct (with regards to a test item)
   - change the instructional medium

RESULTS AND DISCUSSION

The second research question (i.e., does the debriefing procedure produce unique feedback data?) could be related to cost-effectiveness. If debriefing does not provide useful information then it could be argued that developmental testing data might be sufficient. It is important to note that debriefing can be viewed as a procedure from which the results of a testing session are finalized and even if the feedback were of a confirmatory nature, i.e., redundant with the outcomes of developmental testing, then this in itself could be considered as useful information. The analysis conducted for the second question was based on collapsing (ignoring) experimental conditions, pooling responses across units 1 and 2 for the 72 subjects (total N) who participated in the study. All problems identified and explained with their accompanying revisions based on debriefing outcomes from all sources were coded as being either redundant/verified or not with the outcomes of developmental testing.

Results indicate that for problems identified and explained during developmental testing (Source #1) 53.6% were redundant/verified with debriefing outcomes based on retrospective problem explication. Revisions generated from developmental testing (Source #1) were verified/ redundant with debriefing outcomes based on retrospective revision 49.14% of the time. When it came to both take and criticize the test or content related to it during developmental testing (Source #2), only 3.43% of the problem explications made during developmental testing were verified/redundant with the problem explications that test subjects made during debriefing based on posttest error data. With regards to revision, only 2.94% of the revisions made during

developmental testing were verified/redundant with the revisions that test subjects made during debriefing based on posttest error data. For Source #3, the attitudinal survey, 42.56% of the problem explications made during developmental testing were verified/redundant with the problem explications that test subjects made during debriefing based on the results of the survey. For revision, 39.48% of the revisions made during developmental testing were verified/redundant with the revisions that test subjects made during debriefing based on the results of the survey. A conclusion that can be reached based on the descriptive data from this study is that the debriefing procedure did provide a considerable amount of unique information.

Does the test subject characteristic of having indicated prior knowledge of the material affect the overall amount of comments being generated during developmental testing? This is the final question being addressed within this paper. Prior knowledge was not assessed by a pretest measure but was based on test subject self-report (either yes or no) after having completed the second unit. The following question was embedded just after the posttest within the test subjects' "package" of materials: In the past, have you ever been exposed to subject matter content related to the topics of sterilization and disinfection in Microbiology? Thirty-four subjects indicated that they had prior knowledge; 37 subjects indicated that they had no prior knowledge; there was missing data for one subject. All T-test results are based on collapsing over the independent variables. T-tests between the two groups indicated that those who indicated that they had prior knowledge were significantly older and had had more years of schooling than the group who indicated that they did not have prior knowledge. Using prior knowledge (yes or no) as the independent variable, t-tests were applied to the number of comments (problems identified) for unit 1, unit 2, and for the total number of comments. Two-tailed probability values are reported. For unit 1, the t-test did not reveal a significant difference between the two groups ($t = -1.32$, $df = 69$, $p > .05$). For unit 2, the t-test did reveal a significant difference between the two groups ($t = -2.29$, $df = 69$, $p < .05$). Those indicating prior knowledge produced more comments than those who did not indicate prior knowledge. For the two units combined, the t-test did reveal a

significant difference between the two groups ($t = -2.05$, $df = 69$, $p < .05$). Again, those indicating prior knowledge commented more frequently than those not indicating prior knowledge.

CONCLUSION

Although most of the results reported are based on the amounts of data obtained under the experimental conditions for the different sources of feedback, the main focus of the research is the qualitative differences between conditions on the different kinds of feedback obtained using the differenct sources for gathering data. Although there are basically few significant differences on the amounts of data obtained there might well be differences in kind. The analyses necessary to address this important issue have not been completed and will be reported at a later date.

It is useful to note that prior knowledge, based on the results reported here, might be an important variable to consider for future research on the formative evaluation of instructional material. With regards to selecting test subjects, the characteristics that profile optimal test subjects has been a long and frequent debate within the literature. The program evaluation literature often recommends gathering feedback data from graduates of programs. Perhaps quasi-expert review techniques hold promise in future formative evaluation research.

References

Abedor, A. (1971). Development and validation of a model explicating the formative evaluation process for multi-media self-instructional learning systems (Doctoral dissertation, Michigan State University, Lansing, MI, 1971). Dissertation Abstracts International, 32, 4915A.

Andrews, D., & Goodson, L. (1980). A comparative analysis of models of instructional design. Journal of Instructional Development. 3(4), 2 - 16.

Baghdadi, A. (1980). A comparison between two formative evaluation methods (Doctoral dissertation, Indiana University, 1980). Dissertation Abstracts International, 41, 3387A.

Baker, E., & Alkin, M. (1973). Formative evaluation in instructional development. AV Communication Review, 21(4), 389 - 418.

Berthelot, S. (1978). Le niveau de verbalisation d'un sujet dans le cadre de l'application d'une evaluation LVR d'un document audiovisuel par le mode tutoriel. Unpublished doctoral dissertation. Universite de Montreal.

Burkholder, B. (1982). The effectiveness of using the instructional strategy diagnostic profile to prescribe improvements in self-instructional materials teaching abstract concepts. Journal of Instructional Development, 5(2), 2 - 9.

Burt, C., & Geis, G. (1986, April). Guidelines for developmental testing: Proposed and practiced. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Corey, G., & Corey, M. (1982). Groups: Process and practice (2nd ed.). Belmont, CA.: Wadsworth.

Debert, R. (1979). Revisionary tactics. National Society for Performance and Instruction Journal, 5(3), 189 - 196.

Dental Auxiliary Education Project. (1982). Microbiology. New York: Teachers College Press, Columbia University.

Dick, W. (1980). Formative evaluation in instructional development. Journal of Instructional Development, 3(3), 3 - 60.

Dick, W., & Carey, L. (1978). The systematic design of instruction. Glenview, Il.: Scott, Foresman & Company.

Dupont, D. (1980). Developpement et evaluation d'un modele de revision de documents audio-scripto-visuels destine aux reviseurs. GRERDAVE, Rapport 17. Montreal, Quebec: Universite de Montreal.

Dupont, D. & Stolovitch, H. (1983, March). The effects of a systematic revision model on revisers in terms of student outcomes. National Society for Performance and Instruction Journal, 33 - 37.

Frase, L., DeGracie, J., & Poston, W. (1974). Product validation: Pilot test or panel review? Educational Technology, 14(8), 32 - 35.

Geis, G., Burt, C., & Weston, C. (1984, April). Instructional development: Developmental testing. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, (ERIC ED 243 793).

Gropper, G. (1975). Diagnosis and revision in the development of instructional materials. Englewood Cliffs, NJ: Educational Technology Publications.

Henderson. E., & Nathenson, M. (1977). Case study in the implementation of innovation: A new model for developmental testing. In Philip Hills and John Gilbert (Eds.) Aspects of educational technology: Volume X The spread of educational technology. London: Kogan Page.

Horn, R. (1966). Developmental testing. Ann Arbor, MI: Center for Programmed Learning for Business.

Horn, R. (1976). How to write information mapping. Lexington, MA.: Information Resources Inc., P. O. Box 417.

Kandeswamy, S., Stolovitch, H., & Thiagarajan, S. (1976). Learner verification and revision: An experimental comparison of two methods. Audiovisual Communication Review, 24(3). 316 - 328.

McGrath, J., & Altman, I. (1966). Small group research: A synthesis and critique of the field. New York: Holt, Rinehart, & Winston.

Montague, W., Ellis, J., & Wulfeck, W. (1983). Instructional quality inventory: A formative evaluation tool for instructional development. Performance and Instruction Journal, 22(5), 11 - 13.

Newman, S. (1957). Student vs. instructor design of study method. Journal of Educational Psychology, 48(6), 328 - 333.

Robeck, M. (1965). A study of the revision process in programmed instruction. Unpublished master's thesis, University of California, Los Angeles.

Rodriguez, A., & Rodriguez, C. (1982). Application de l'approche L.V.R. a differents stades de la realisation de materiel d'enseignement audio-visuel. Unpublished master's thesis, University of Montreal.

Rosen, M. (1968). Final report: An experimental design for comparing the effects of instructional media programing procedures. Palo Alto, California: American Institutes for Research in Behavioral Sciences. Sponsored by the U. S. Department of Health, Education, and Welfare, Project No. 5 - 8332.

Rothkopf, E. (1963). Some observations on predicting instructional effectiveness by simple inspection. The Journal of Programmed Instruction, 2, 19 - 20.

Smith, K. (1978). The utilization of teacher feedback in ETV development. Educational Technology, 18(19), 49 - 51.

Stolovitch, H. (1982). Applications of the intermediate technology of learner verification and revision (LVR) for adapting international instructional resources to meet local needs. Performance and Instruction Journal, 21(7), 16 - 22.

Wager, J. (1983). One-to-one and small group formative evaluation. Performance and

Instruction Journal, 22(5), 5 - 7.

Weston, C. (1986). Formative evaluation of instructional materials: An overview of approaches. Canadian Journal of Educational Communication, 15(1), 5 - 17.

Weston, C. (1987). The importance of involving experts and learners in formative evaluation. Canadian Journal of Educational Communication, 16(1), 45 - 58.

APPENDIX 1

TABLE 1

Two-Way ANOVA of the Number of Problems Identified (Comments) during Developmental Testing

Unit 1: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	10.125	1	10.125	1.128	.292
Number of Participants (N)	42.014	1	42.014	4.682	.034 *
R x N Interaction	4.014	1	4.014	.447	.506
Residual	610.167	68	8.973		
Total	666.319	71	9.385		

Unit 2: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	13.491	1	13.491	1.006	.319
Number of Participants (N)	11.827	1	11.827	0.882	.351
R x N Interaction	46.419	1	46.419	3.462	.067
Residual	898.281	67	13.407		
Total	969.662	70	13.852		

Total: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	45.825	1	45.825	1.265	.265
Number of Participants (N)	96.267	1	96.267	2.657	.108
R x N Interaction	76.872	1	76.872	2.122	.150
Residual	2427.497	67	36.231		
Total	2644.592	70	37.780		

* $p < .05$

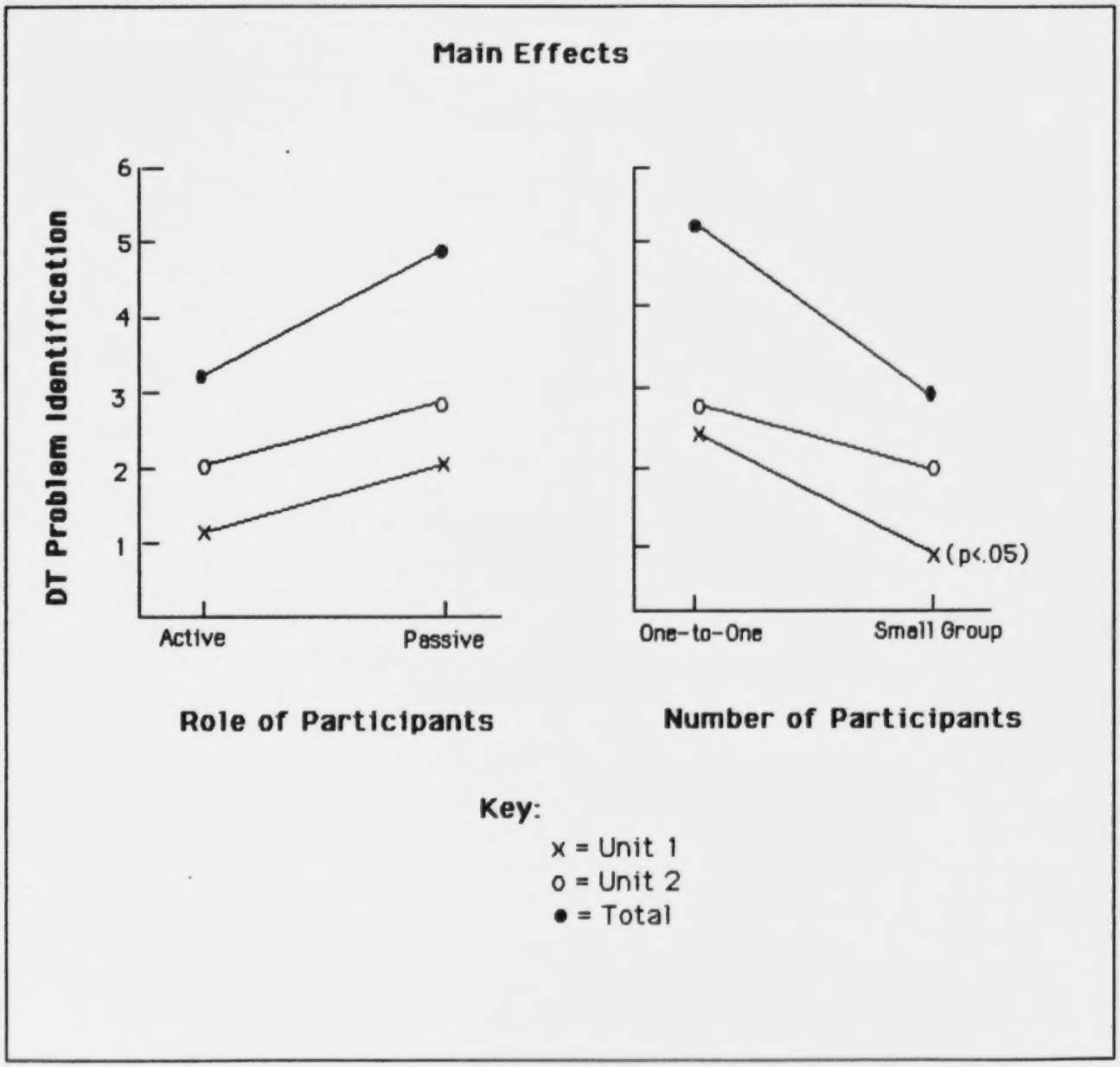

Figure 2. Plotting of means: Main effects of role and number of participants on the number of problems identified by test subjects commenting during the developmental testing of unit 1, unit 2, and total for the instructional material.

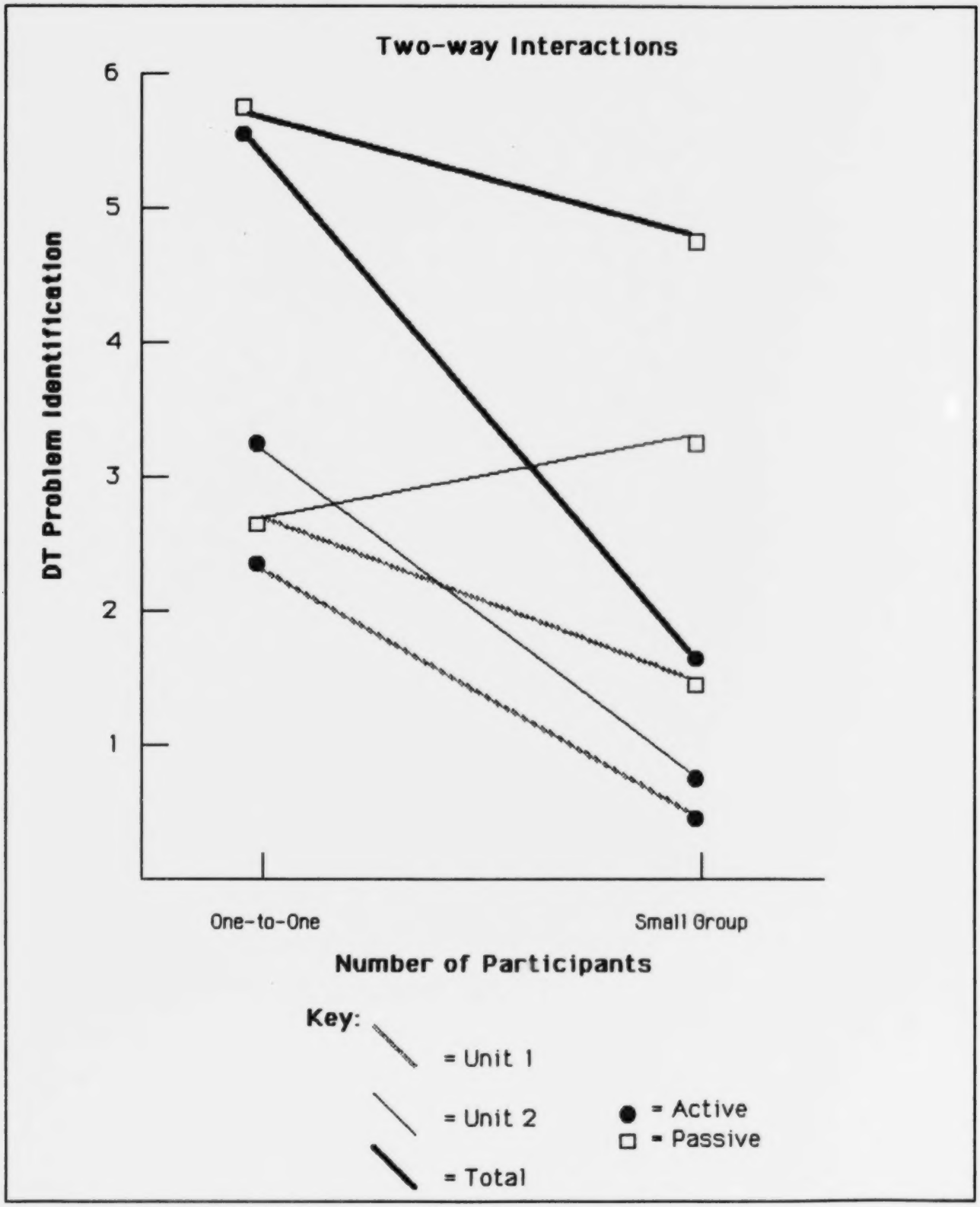

Figure 3. Plotting of means. Interaction between role and number of participants with the number of problems identified by test subjects commenting during the developmental testing of unit 1, unit 2, and total for the instructional material.

TABLE 2
Two-Way ANOVA of the Number of Comments that were Debriefed

Unit 1: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	0.188	1	0.188	0.012	.914
Number of Participants (N)	115.147	1	115.147	7.158	.009**
R x N Interaction	5.475	1	5.475	.340	.562
Residual	1077.778	68	16.086		
Total	1198.479	71	17.121		

Unit 2: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	21.905	1	21.905	1.855	.178
Number of Participants (N)	0.000	1	0.000	0.000	1.000
R x N Interaction	22.549	1	22.549	1.910	.172
Residual	791.007	67	11.806		
Total	835.465	70	11.935		

Total: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	37.556	1	37.556	1.012	.318
Number of Participants (N)	122.722	1	122.722	3.308	.073
R x N Interaction	43.556	1	43.556	1.174	.282
Residual	2522.778	68	37.100		
Total	2726.611	71	38.403		

** $p < .01$

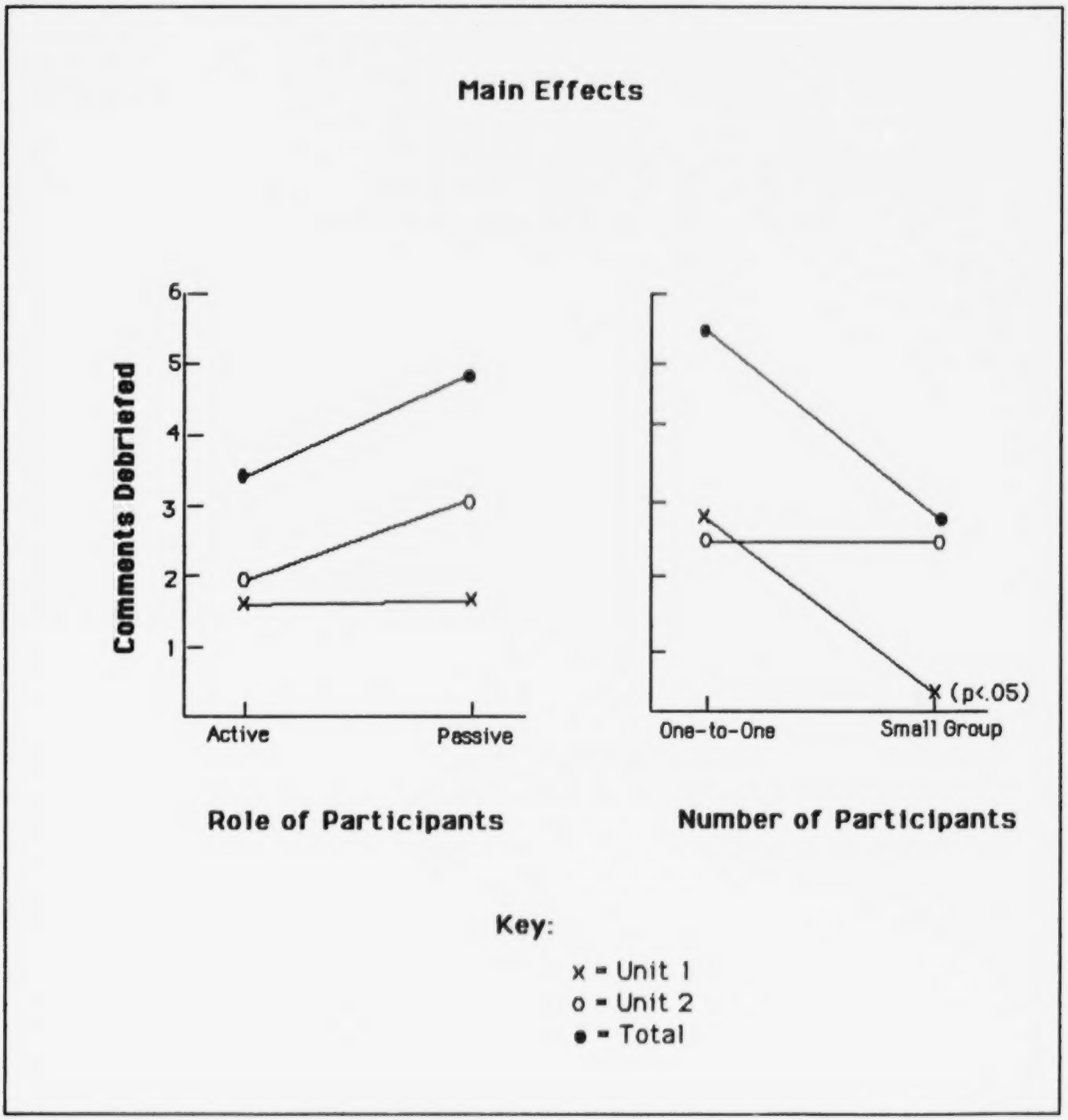

Figure 4. Plotting of means: Main effects of role and number of participants on the number of debriefing agenda items based on comments for unit 1, unit 2, and total for the instructional material.

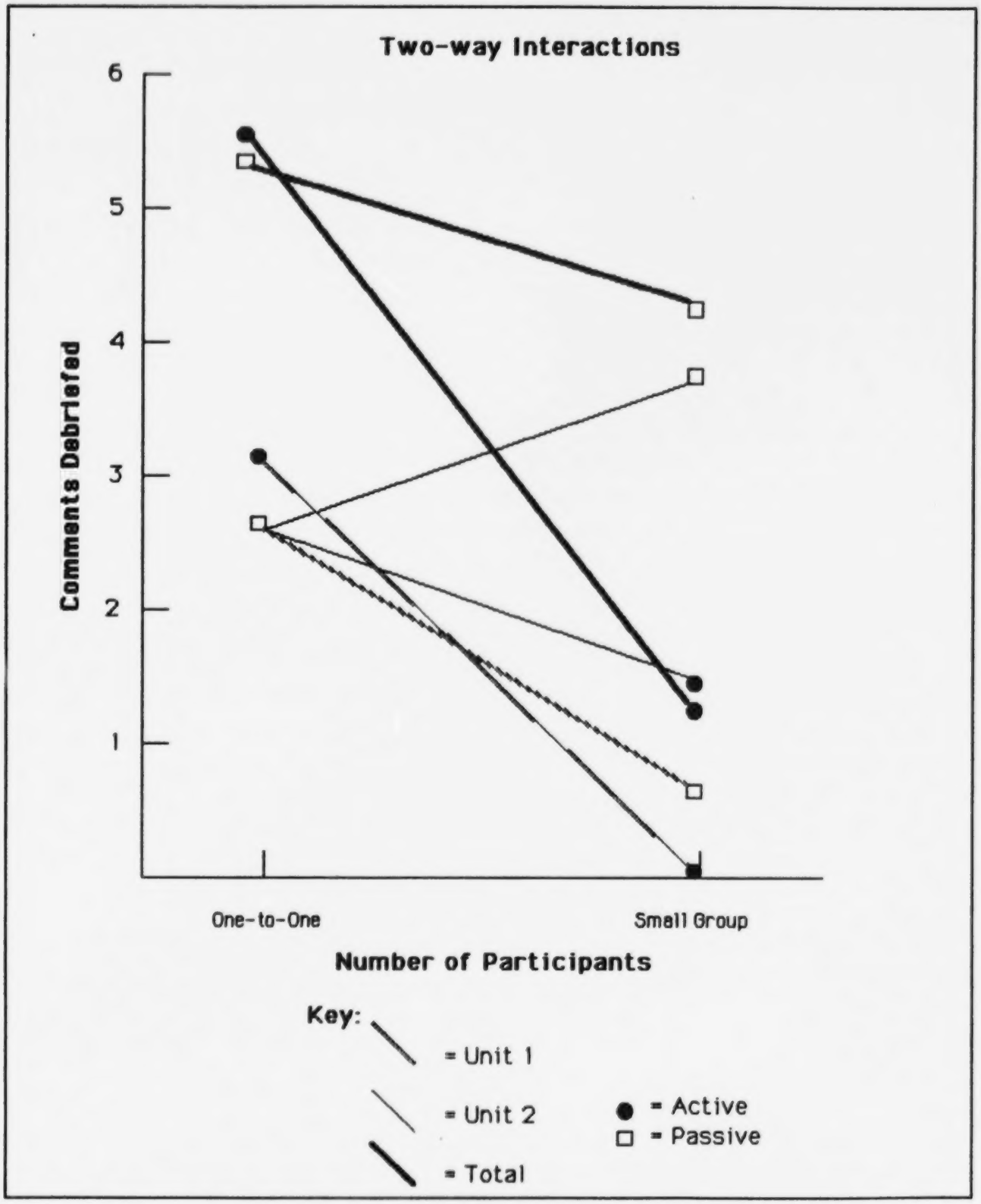

Figure 5. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on comments for unit 1, unit 2, and total for the instructional material.

TABLE 3

Two-Way ANOVA of the Number of Debriefing Agenda Items Based on Posttest Errors

Unit 1: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	3.125	1	3.125	3.329	.072
Number of Participants (N)	17.014	1	17.014	18.124	.0001****
R x N Interaction	11.681	1	11.681	12.443	.001 ***
Residual	63.833	68	.939		
Total	95.653	71	1.347		

Unit 2: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	0.415	1	0.415	0.182	.671
Number of Participants (N)	32.194	1	32.194	14.079	.0001****
R x N Interaction	9.169	1	9.169	4.010	.049 *
Residual	153.209	67	2.287		
Total	195.099	70	2.787		

Total: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	8.681	1	8.681	2.024	.159
Number of Participants (N)	86.681	1	86.681	20.213	.0001****
R x N Interaction	48.347	1	48.347	11.274	.001 ***
Residual	291.611	68	4.288		
Total	435.319	71	6.131		

* $p < .05$
*** $p < .001$
**** $p < .0001$

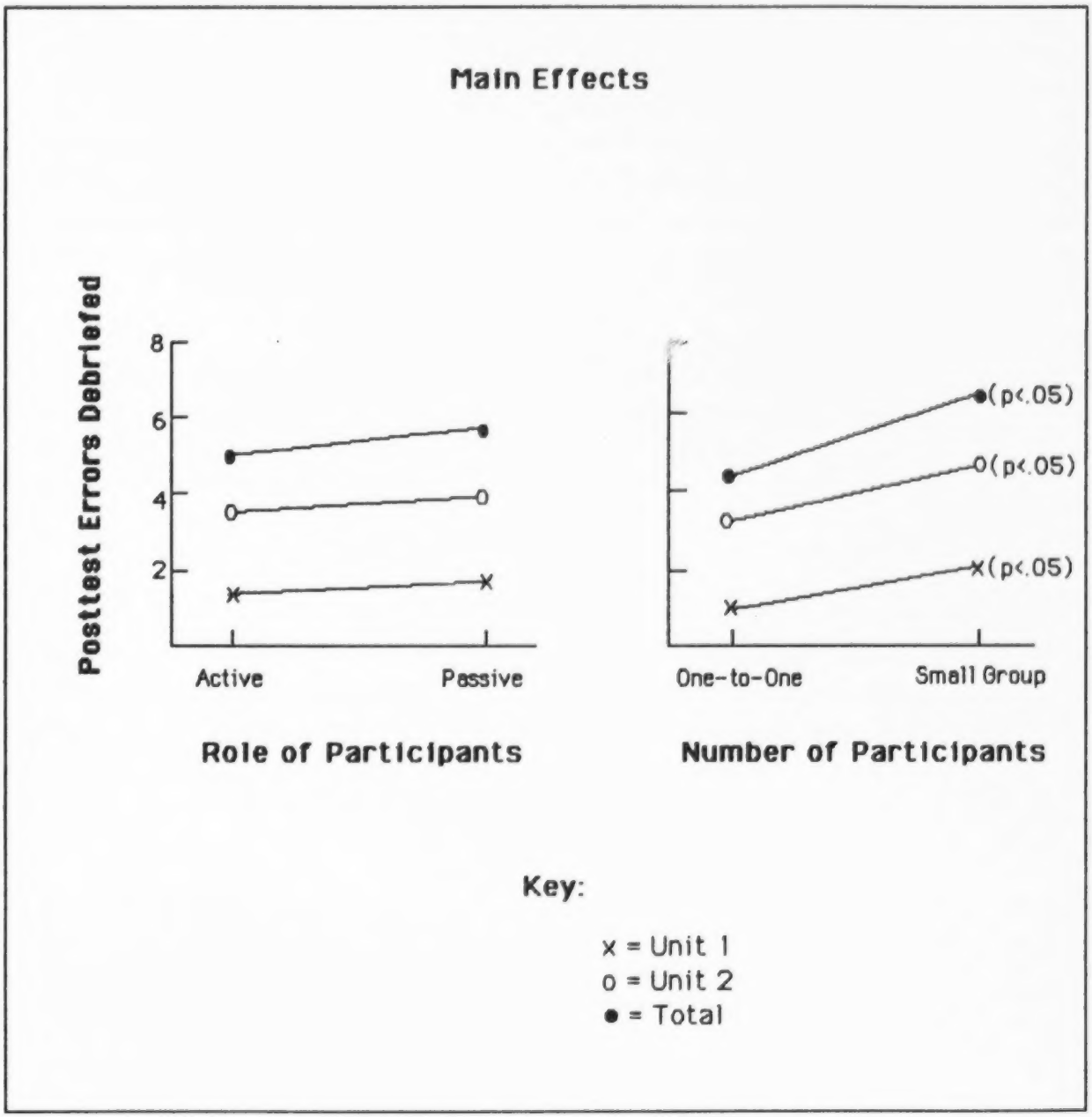

Figure 6. Plotting of means: Main effects of role and number of participants on the number of debriefing agenda items based on posttest errors for unit 1, unit 2, and total for the instructional material.

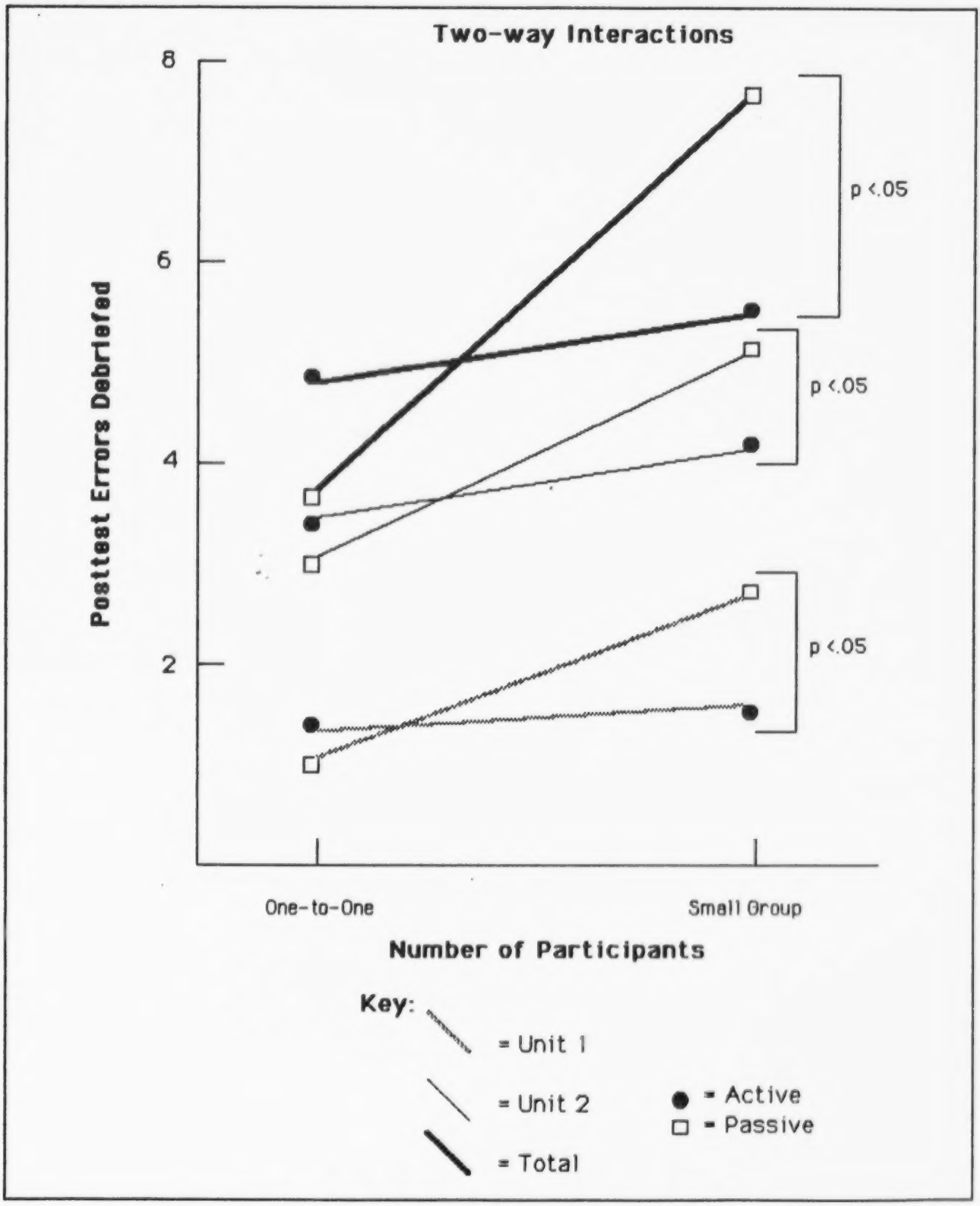

Figure 7. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on posttest errors for unit 1, unit 2, and total for the instructional material.

TABLE 4

Two-Way ANOVA of the Number of Debriefing Agenda Items Based on the Attitudinal Survey Responses

Unit 1: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	18.000	1	18.000	0.897	.347
Number of Participants (N)	2.000	1	2.000	.100	.753
R x N Interaction	43.556	1	43.556	2.170	.145
Residual	1364.889	68	20.072		
Total	1428.444	71	20.119		

Unit 2: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	136.746	1	136.746	7.258	.009**
Number of Participants (N)	148.241	1	148.241	7.869	.007**
R x N Interaction	50.637	1	50.637	2.688	.106
Residual	1262.248	67	18.840		
Total	1602.000	70	22.886		

Total: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	19.014	1	19.014	.457	.501
Number of Participants (N)	268.347	1	268.347	6.446	.013*
R x N Interaction	30.681	1	30.681	.737	.394
Residual	2830.944	68	41.632		
Total	3148.986	71	44.352		

* $p < .05$
** $p < .01$

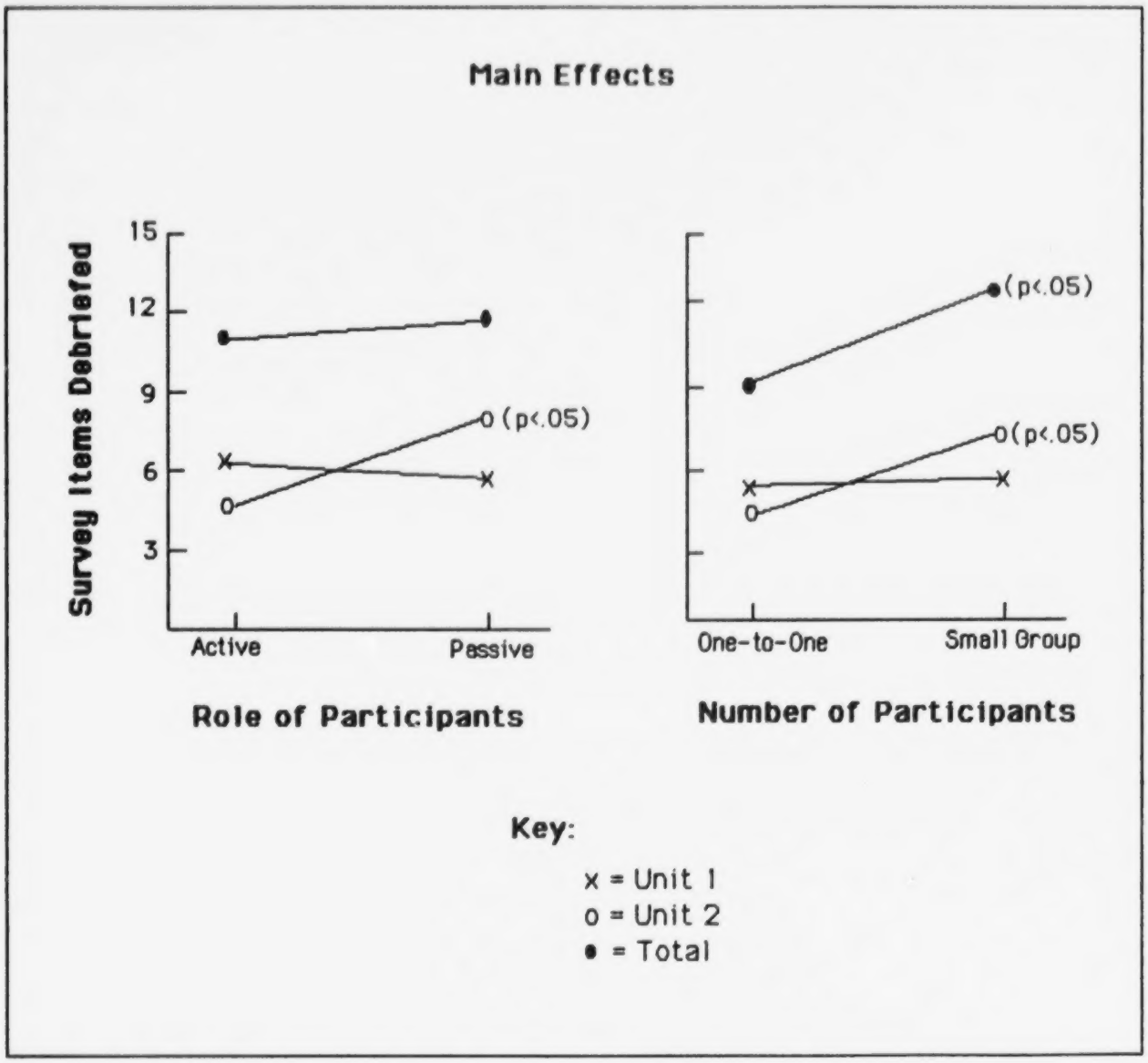

Figure 8. Plotting of means: Main effects of role and number of participants on the number of debriefing agenda items based on survey responses to unit 1, unit 2, and total for the instructional material.

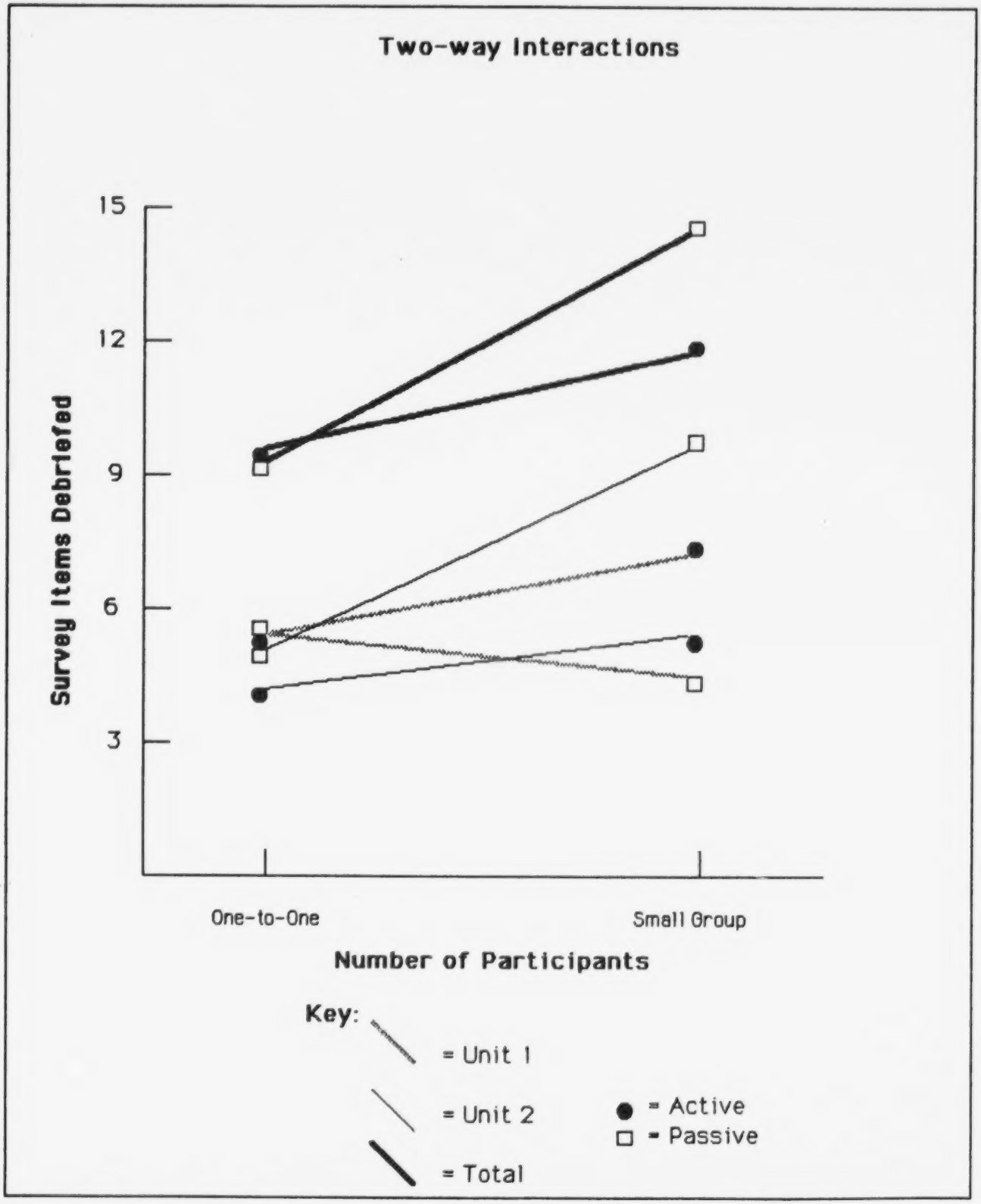

Figure 9. Plotting of means. Interaction between role and number of participants with the number of debriefing agenda items based on survey responses to unit 1, unit 2, and total for the instructional material.

TABLE 5

Two-Way ANOVA of the Total Number of Debriefing Agenda Items from all Sources Combined

TOTAL: Source	Sum of squares	*df*	Mean squares	F	Significance of F
Role of Participants (R)	150.222	1	150.222	1.218	.274
Number of Participants (N)	220.500	1	220.500	1.787	.186
R x N Interaction	304.222	1	304.222	2.466	.121
Residual	8388.556	68	123.361		
Total	9063.500	71	127.655		

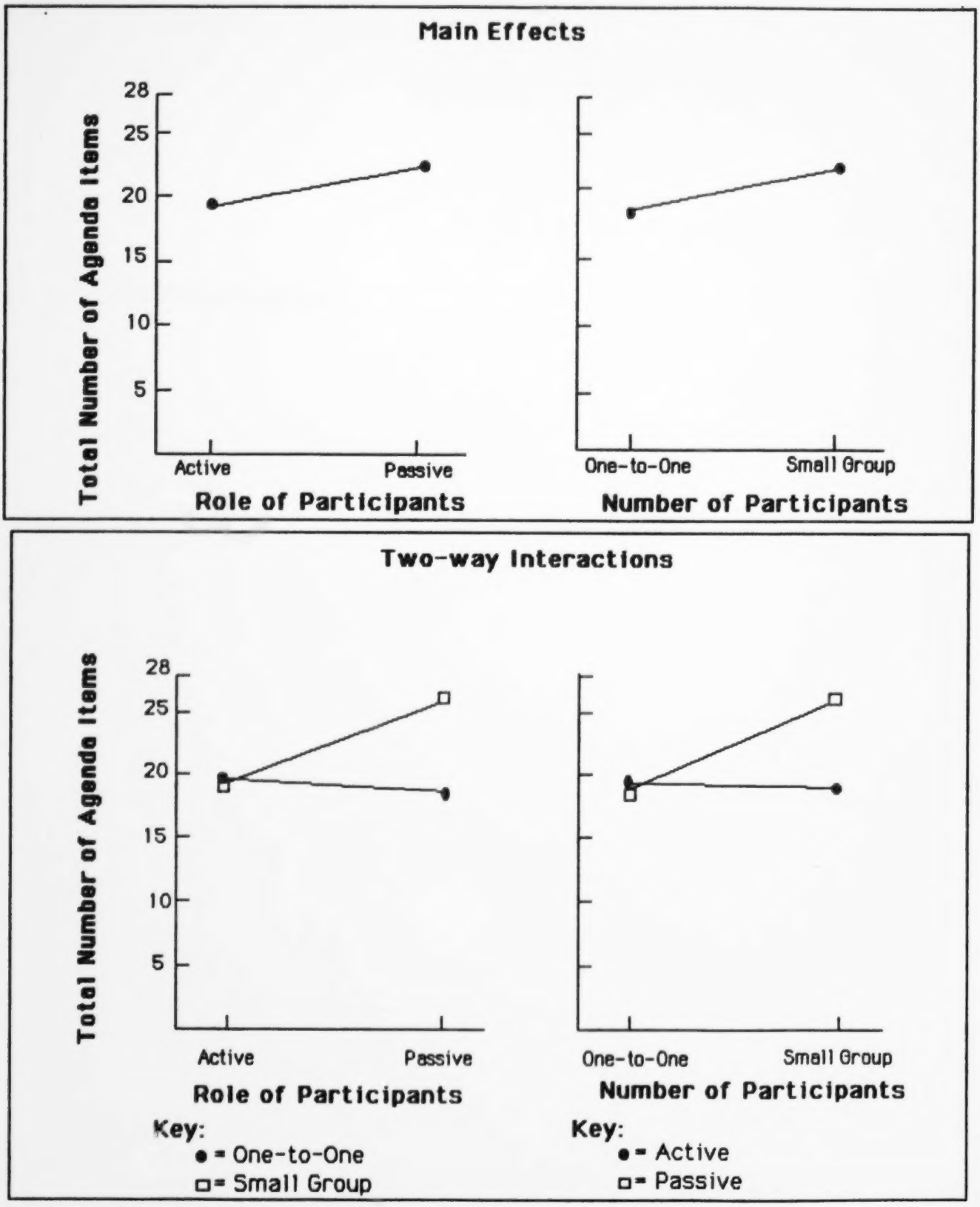

Figure 10. Plotting of means: Main effects of role and number of participants and their interactions with the total number of agenda items (from all feedback sources) debriefed for both units of instructional material.